Melbourne Bioinformatics

BIOINFORMATICS + DATA SERVICES + INFRASTRUCTURE, FOR LIFE SCIENCES TODAY

https://www.melbournebioinformatics.org.au/tutorials/tutorials/qiime2/qiime2/

ARDC Nectar Research Cloud

https://dashboard.rc.nectar.org.au/project/

https://qiime2.org/

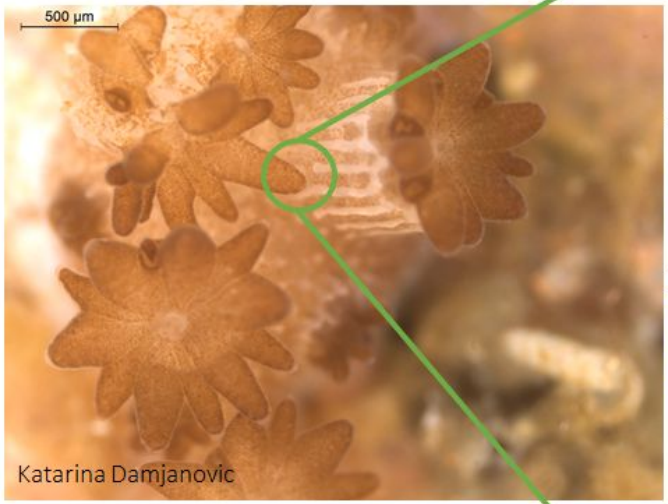
https://view.qiime2.org/

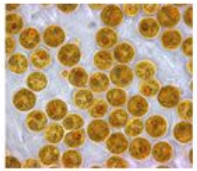
Linux/Unix/macOS command line

- Tab: autofill (if it doesn’t autofill something is incorrect)
- Ctrl-C: Abort command
- ls: list directory contents
- tree: visualize directories, recursively
- pwd: print working (i.e., current) directory
- cd: change directory
- mkdir: make directory
- rmdir: remove a directory
- nano: open a text editor
- cp: copy a directory or a file
- cat/more/less: print contents of a file to the terminal
- rm: remove a file (rm -r: removes a directory)
- mv: move (i.e., rename) a directory or a file
- head: print the first ten lines of a file to the terminal
- tail: print the last ten lines of a file to the terminal
- curl or wget: download a file from a URL (you will see this in other QIIME2 tutorials)
- man: learn about a command (also, most other cmds: -h; --help)

Cnidarian **holobiont**

Coral

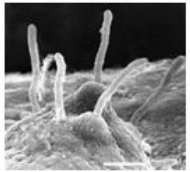




Symbiodiniaceae

Bacteria, Archaea

Viruses

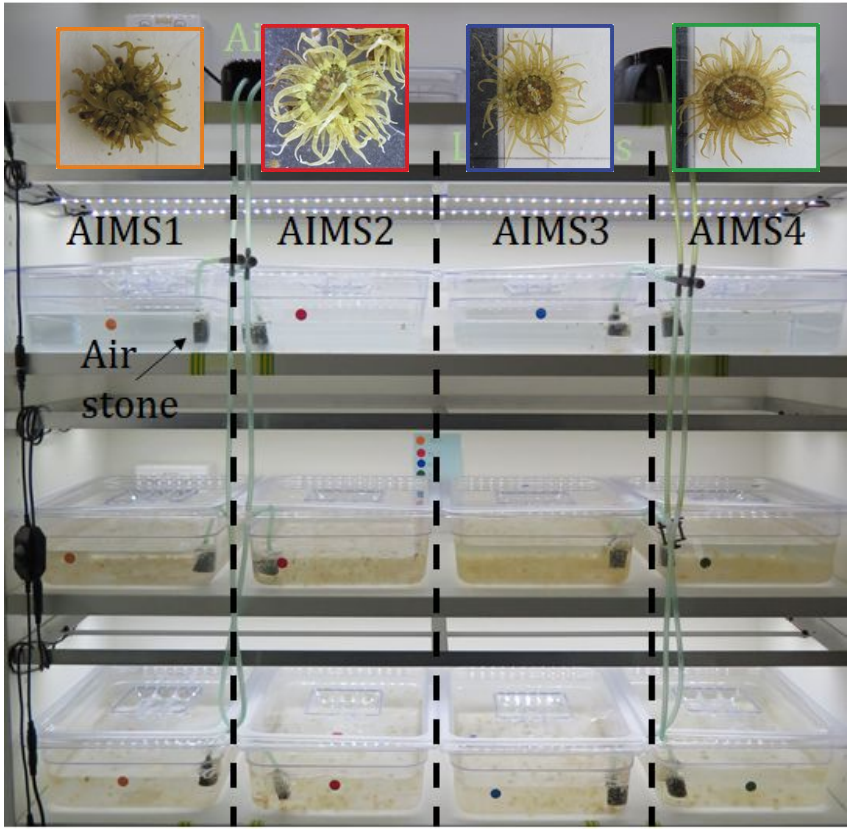


Fungi

Exaiptasia diaphana

Rohwer et al., 2002; Ricci et al., 2019

Background on data



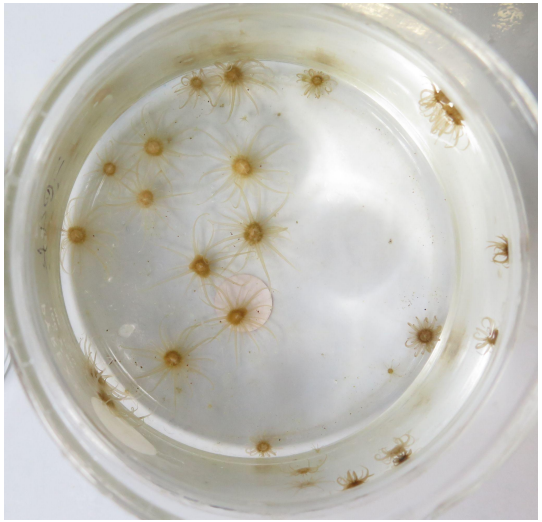
Short-Term Exposure to Sterile Seawater Reduces Bacterial Community Diversity in the Sea Anemone, *Exaiptasia diaphana*

Ashley M. Dungan[1], Madeleine J. H. van Oppen[1,2] and Linda L. Blackall[1]*

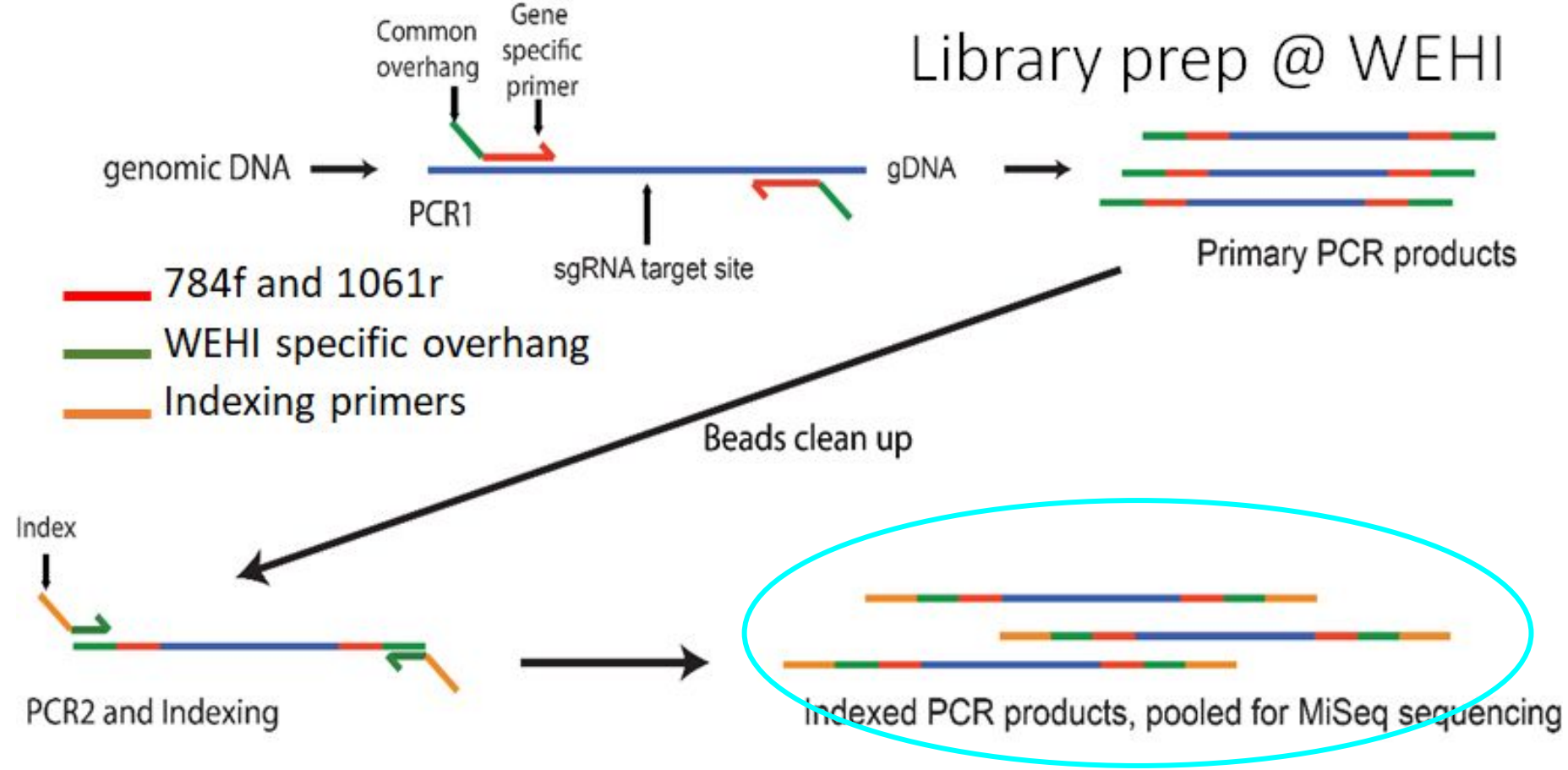
[1] School of BioSciences, The University of Melbourne, Melbourne, VIC, Australia, [2] Australian Institute of Marine Science, Townsville, QLD, Australia

frontiers in Marine Science

Sterile SW
3 weeks



Library prep @ WEHI



Adapted from Aubrey et al., 2014

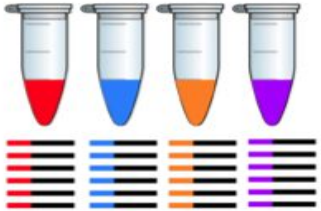
Import data into QIIME2

What do you know about your data?

- Single vs paired end?
  - Single: one direction of sequencing
  - Paired: forward and reverse reads
- Multiplexed vs demultiplexed?
  - Multiplexed: fastq.gz file(s) for each read set and another that contains the associated barcodes
  - Demultiplexed: one fastq.gz file per sample

Multiplexed Data

Barcoded per-sample



Pool and sequence samples

Track per-sample barcodes (e.g., in spreadsheet)

sample-metadata.tsv

SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

sequences.fastq(.gz)

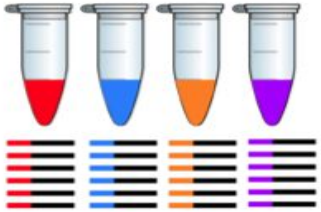
```
@HWI-6X_9267:1:1:25:1051
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTG
GCTTGGTAAGTCCATGGTGAAATCCCTCGGCTCAACCGAGGAACTG
+
abaaaaa`^`a_]^\``\``a`^`]]]^^`a[VXGX``Z_\\\_^\a^SYOZVV
SVYGYVDXOZVT\TITBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI-6X_9267:1:1:25:267
TACGTATGGGGCAAGCGTTATCCGGAATTATTGCGCGTAAAGACTGCGTAGGTG
GTGGCTTAAGCGCAGGGTTTAAGGCAAT
+
aa^^[__^^^_^_^[^^[^^____ZZ[^
XUWWURZUYY]XXRZRNVTRTNTWUUU^
@HWI-6X_9267:1:1:25:609
TACGTAGGGGGCAAGCGTTATCCGGATT
GATGGACAAGTCTGATGTGAAAGGCTGG
+
aaab`aaa`aaaaaaaaaaaaaaa^aa
[][I^^aZZ^WW^_^`ZZ_T]XY^^\^Z
@HWI-6X_9267:1:1:25:519
GACGGAGGATGCAAGTGTTATCCGGAAT
GTTTACTAAGTCAACTGTTAAATCTTGA
+
abaaaaaa`aaaaaa\aaaaaaa```aa
WY]]_Z_XX\\[]]]^`[\XTVX]`T_V
@HWI-6X_9267:1:1:25:1109
TACGGAGGGTGCGAGCGTTAATCGGAAT
GTTAGGTAAGTCAGATGTGAAAGCCCCG
+
aaaba^`a^N_`\_``a_a]Zaa^^\Z`
^RVH_PHOWZM[PTRPTRYUBBBBBBBB
```

barcodes.fastq(.gz)

```
@HWI-6X_9267:1:1:25:1051
AACGCAC
+
bbbbbbb
@HWI-6X_9267:1:1:25:267
AAGAGAT
+
bbbbbbb
@HWI-6X_9267:1:1:25:609
AACGCAC
+
bbbbbbb
@HWI-6X_9267:1:1:25:519
ACAGCAG
+
bbbbbbb
@HWI-6X_9267:1:1:25:1109
ACAGCTA
+
bbbbbbb
@HWI-6X_9267:1:1:25:434
ACACGAG
+
```

Demultiplexed Data

Barcoded per-sample



Pool and sequence samples

Track per-sample barcodes (e.g., in spreadsheet)

sample-metadata.tsv

SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

4ac2.fastq(.gz)

e375.fastq(.gz)

4gd8.fastq(.gz)

9872.fastq(.gz)

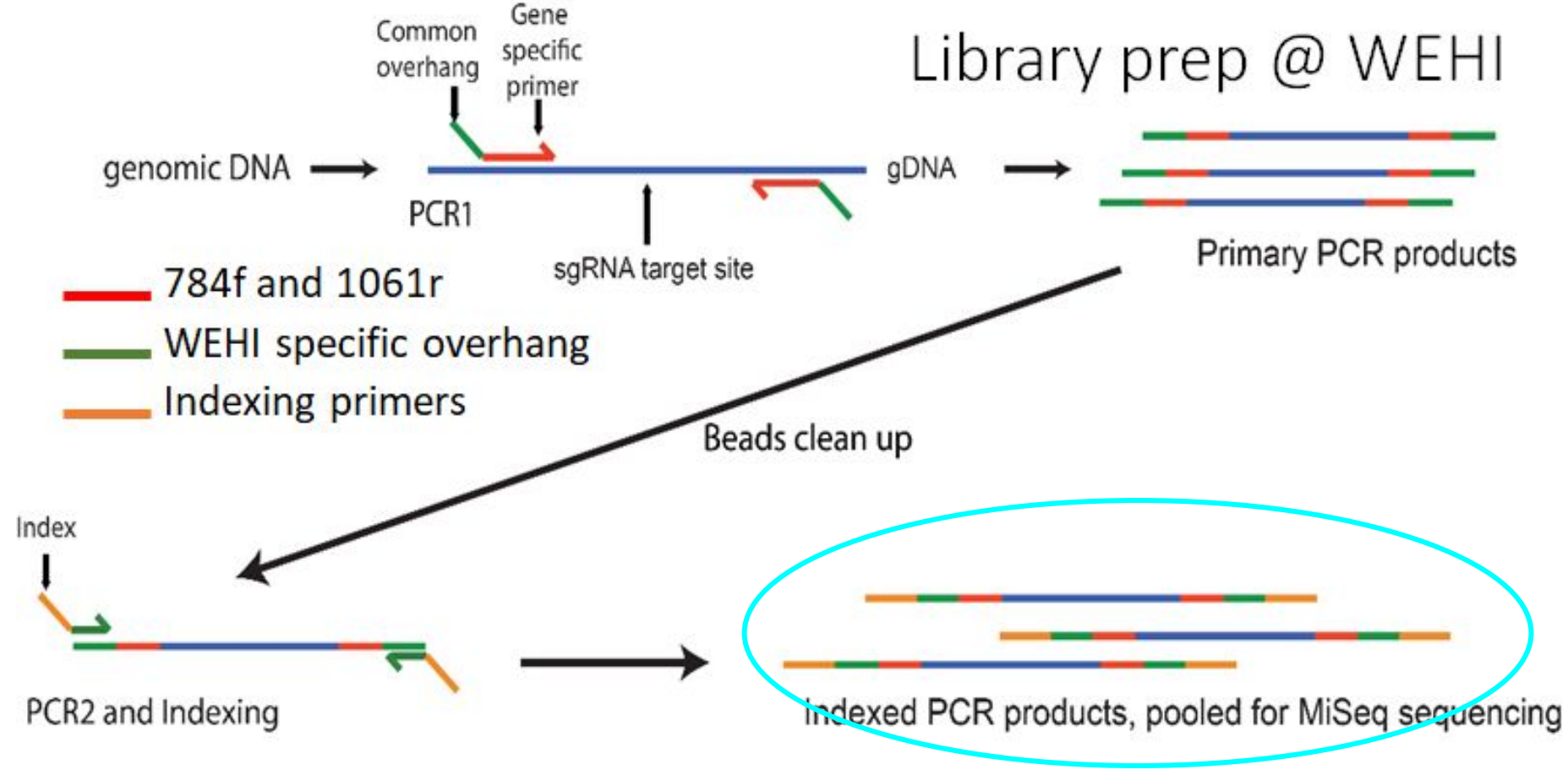
```
@HWI-6X 9267:1:1:25:1109
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAA
AGCGTACGTAGGCGGTTAGGTAAGTCAGATGTGAAAGCCC
CGGGCTCCACCTGGGAATGG
+
aaaba^`a^N `\ ``a a]Zaa^^\Z`[M]a`[VYa^ X
^ Z]NZ\`]TY\] ^RVH PHOWZM[PTRPTRYUBBBBBB
BBBBBBBBBBBBBBBBBBB
```

What do you know about your data?

- Single vs paired end?
  - Single: one direction of sequencing
  - Paired: forward and reverse reads
- Multiplexed vs demultiplexed?
  - Multiplexed: fastq.gz file(s) for each read set and another that contains the associated barcodes
  - Demultiplexed: one fastq.gz file per sample
- Have your adapters and primers been removed?
- Will your files come zipped? (ending in .gz)

Unsure? **Make sure you ask the sequencing facility and know the answers to these specific details.**

Library prep @ WEHI



Adapted from Aubrey et al., 2014

Cutadapt = cutting off adapters (overhang+primer)

```
=== Summary ===

Total read pairs processed:                 13,122
  Read 1 with adapter:                      13,122 (100.0%)
  Read 2 with adapter:                      13,122 (100.0%)
Pairs that were too short:                       0 (0.0%)
Pairs written (passing filters):            13,122 (100.0%)
```

Overview of removed sequences

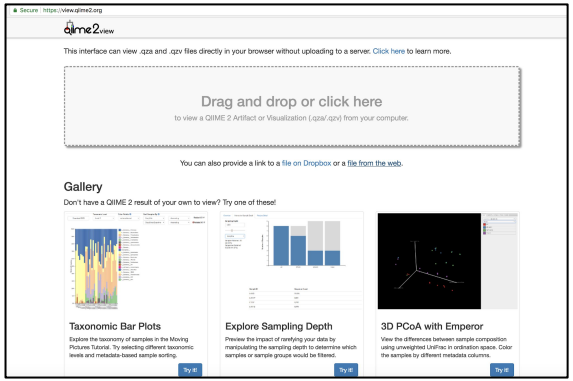
length	count	expect	max.err	error counts
43	1	0.0	3	1
45	1	0.0	3	1
46	19	0.0	3	14 3 0 2
47	106	0.0	3	62 27 17
48	1047	0.0	3	705 330 9 3
49	11931	0.0	3	11512 405 14
50	17	0.0	3	4 12 1

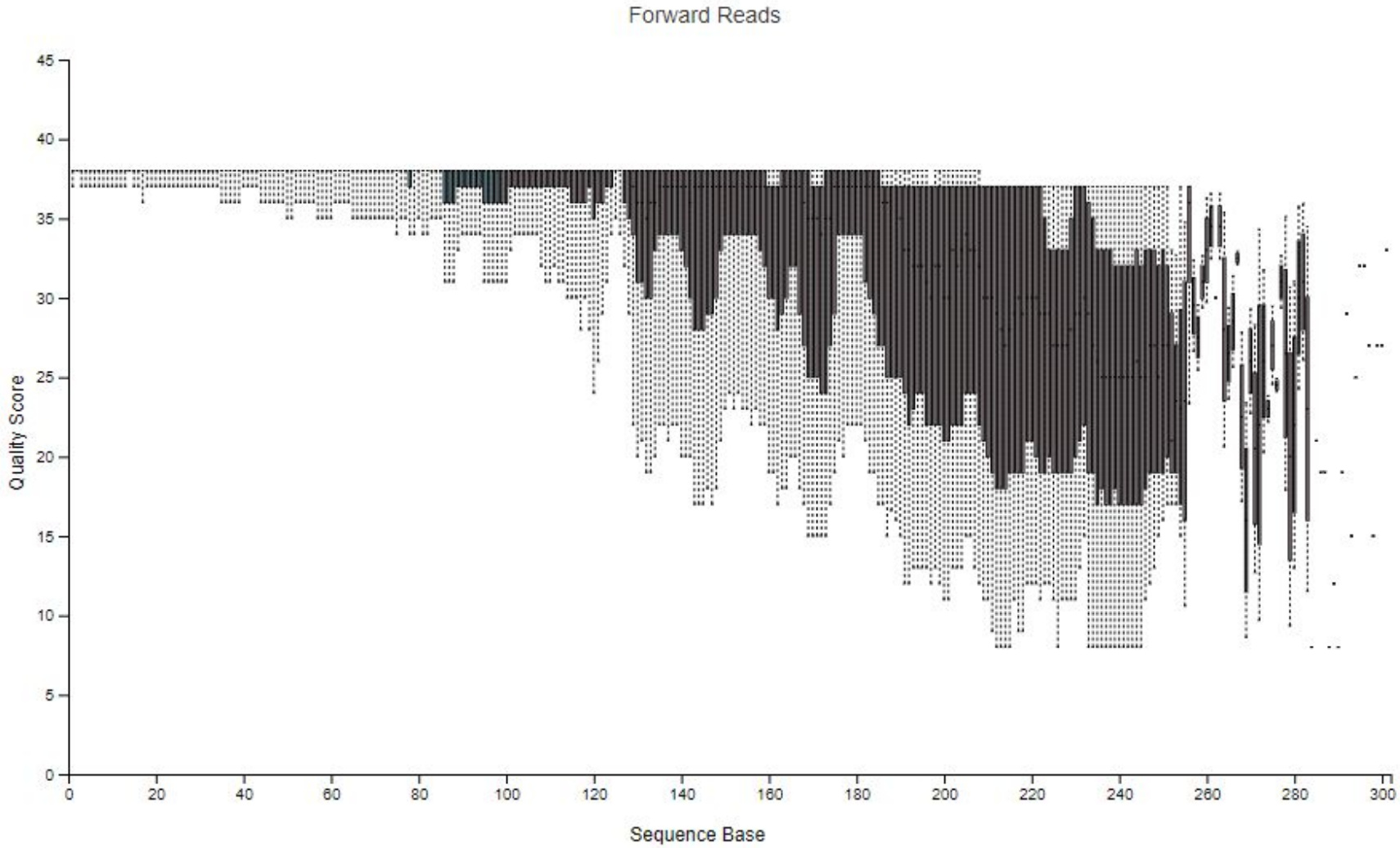
Accessing output files

- Mac users:

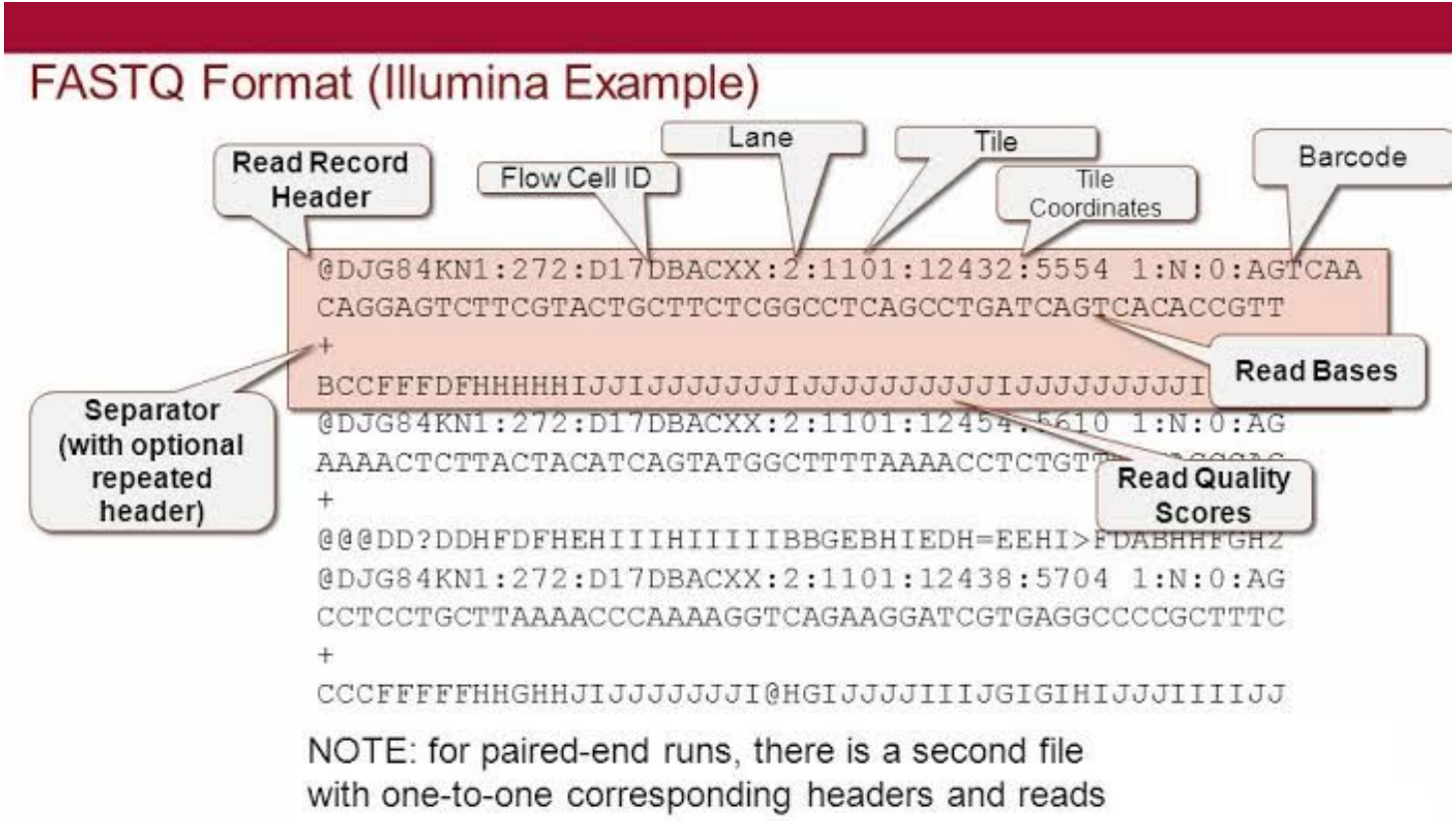
  scp FILENAME username@your_IP_address:/PATH/TO/TARGET/FOLDER/

- Windows users: Use FileZilla to transfer to your local drive
- Go to https://view.qiime2.org/
- Drag file into qiime2 view




Forward Reads
Quality Score
45
40
35
30
25
20
15
10
5
0
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
Sequence Base

Quality Scores



Phred Quality Score = Q-score

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

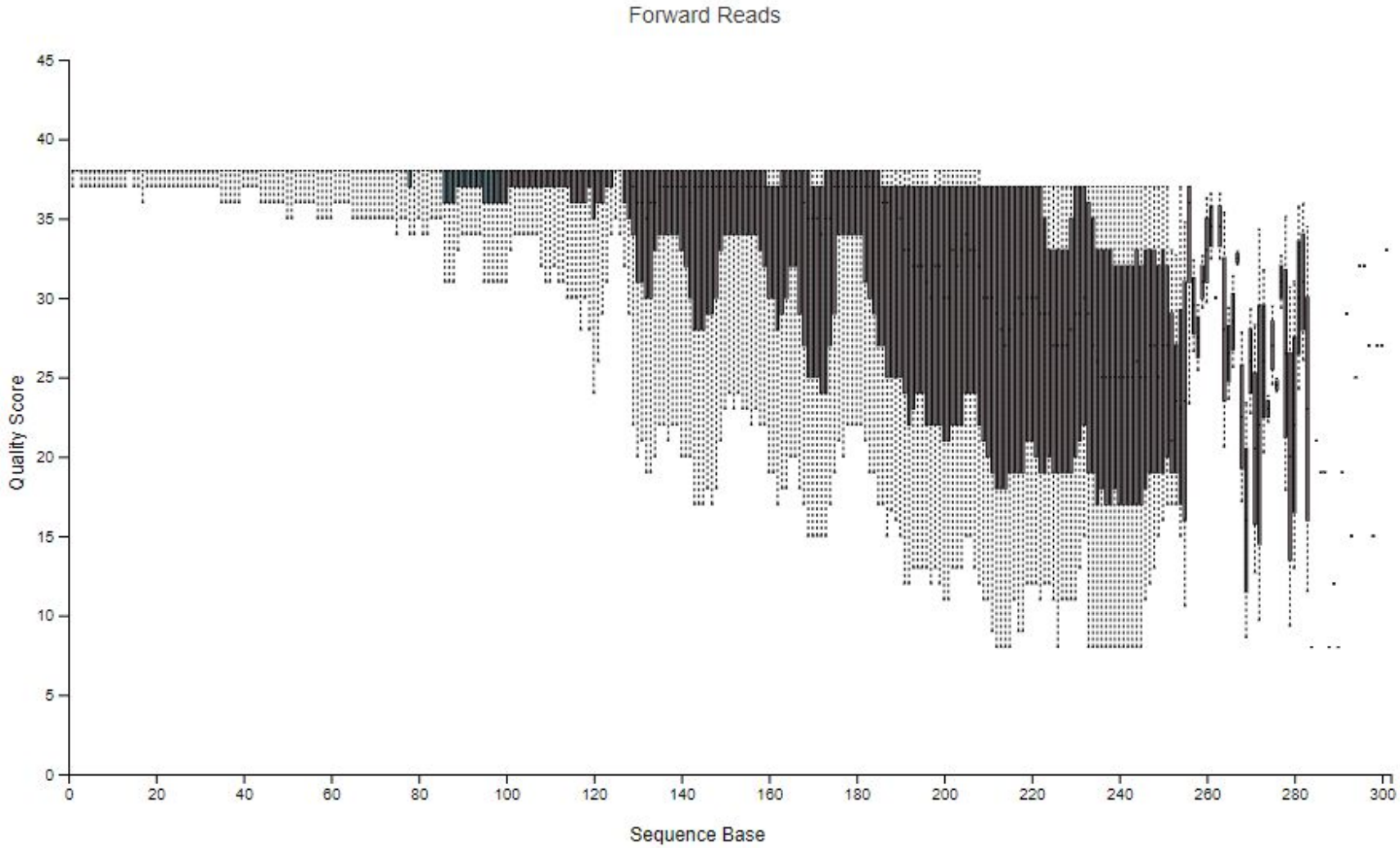
Quality Score Encoding

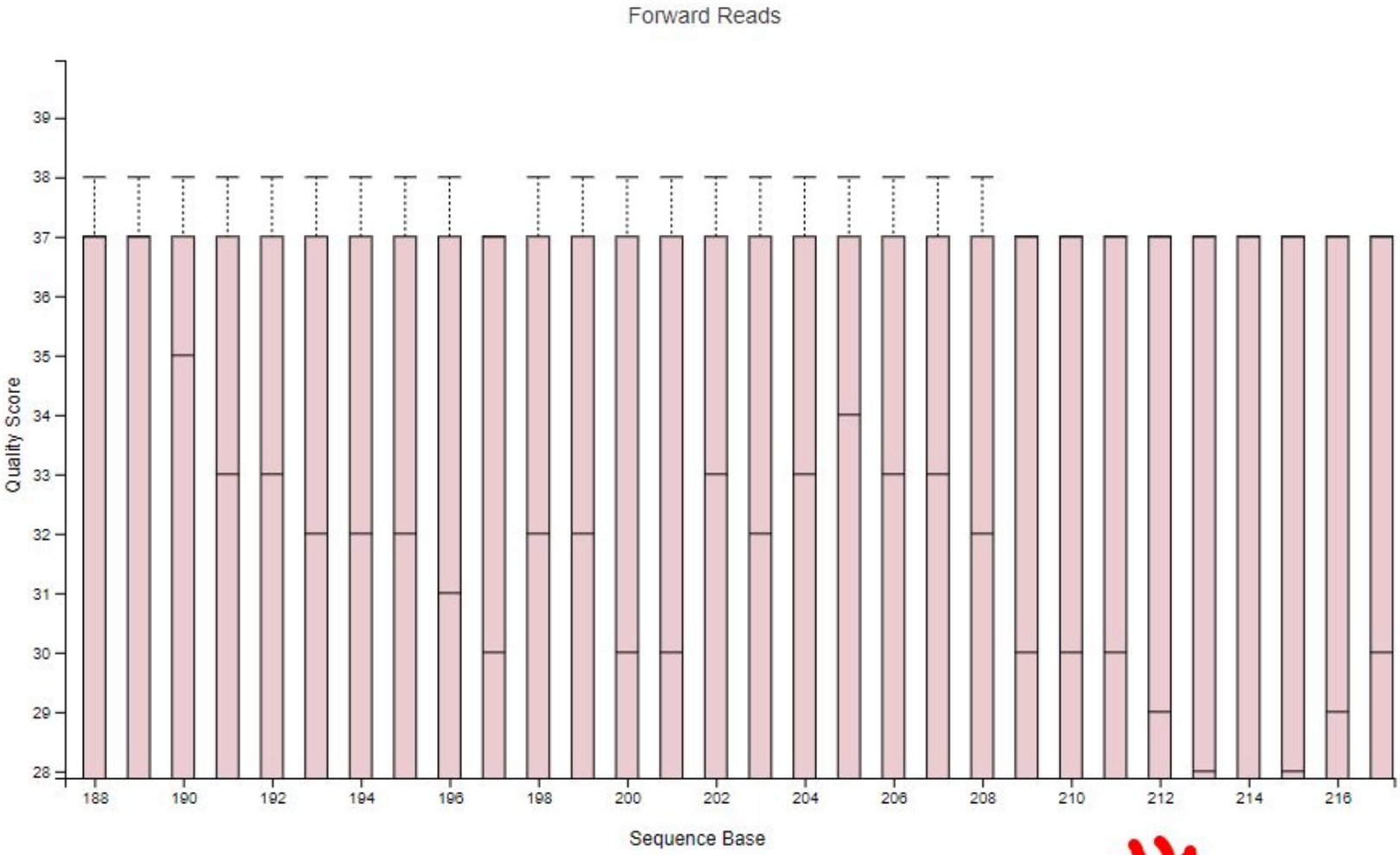
In FASTQ files, quality scores are encoded into a compact form, which uses only 1 byte per quality value. In this encoding, quality score is represented as the character with an ASCII code equal to its value + 33. The following table demonstrates relationship between the encoding character, its ASCII code, and the quality score represented.

NOTE
When Q-score binning is in use, the subset of Q-scores applied by the bins is displayed.

Table 2 ASCII Characters Encoding Q-scores 0–40

Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score
!	33	0	6	54	21
"	34	1	7	55	22
#	35	2	8	56	23
$	36	3	9	57	24
%	37	4	:	58	25
&	38	5	;	59	26
'	39	6	<	60	27
(	40	7	=	61	28
)	41	8	>	62	29
*	42	9	?	63	30
+	43	10	@	64	31
,	44	11	A	65	32
-	45	12	B	66	33
.	46	13	C	67	34
/	47	14	D	68	35
0	48	15	E	69	36
1	49	16	F	70	37
2	50	17	G	71	38
3	51	18	H	72	39
4	52	19	I	73	40
5	53	20			


Forward Reads
Quality Score
45
40
35
30
25
20
15
10
5
0
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
Sequence Base


Forward Reads
Quality Score
39
38
37
36
35
34
33
32
31
30
29
28
188
190
192
194
196
198
200
202
204
206
208
210
212
214
216
Sequence Base

DADA2: What is it?

- **D**ivisive **A**mplicon **D**enoising **A**lgorithm, version **2** (Callahan et al. 2016)
- DADA2 …
  - … is a software package (QIIME2 add-on) that models and corrects Illumina-sequenced amplicon errors
  - … infers sample sequences exactly and resolves differences of as little as one nucleotide (ASVs). This allows for the identification of variants and reveal diversity in a given taxonomic group
  - … is reference free and applicable to any genetic locus

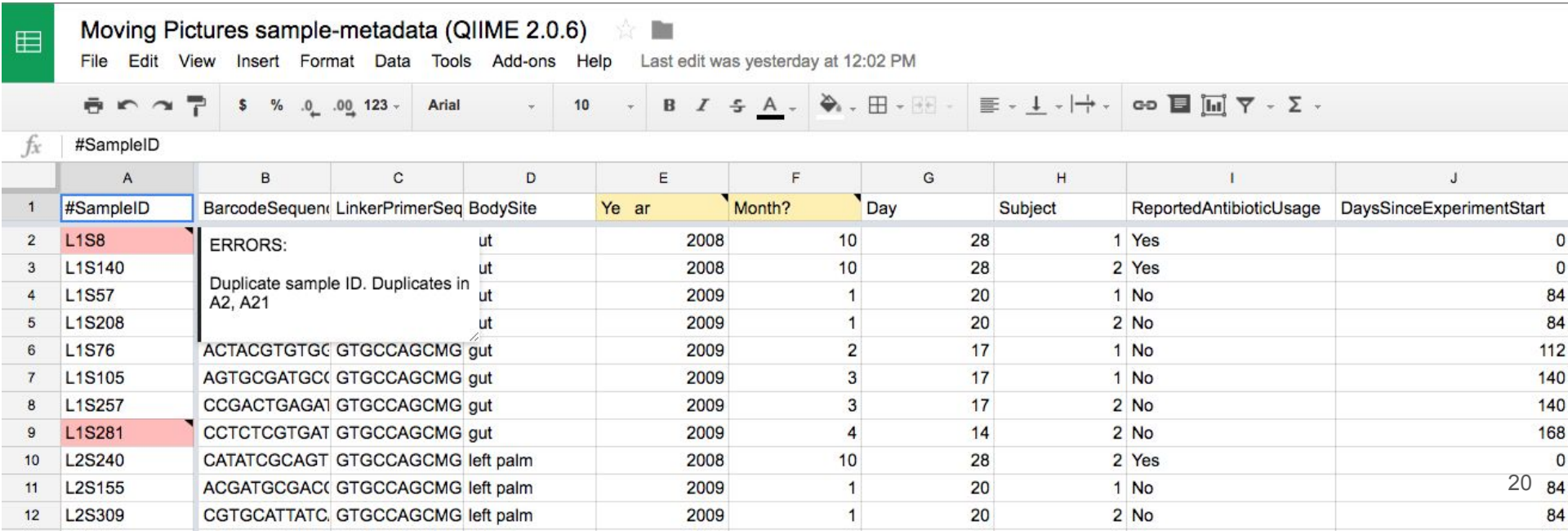
DADA2: How does it do that?

- Denoising
  - Filtering - user defined. Trims sequences to a specified length, removes sequences shorter than that length
  - Model errors within a read and between reads
  - Abundance - sequences too abundant to be explained by errors in sequencing are kept
  - Sequence comparison (i.e. excluding reads whose pairs have >10% mismatch)
- Clustering
  - Reads with exact overlaps are merged by sample
  - Reads with the same sequence are grouped into unique sequences with an associated abundance and consensus quality profile
  - These are called **A**mplicon **S**equencing **V**ariants (ASVs) or Features in some tutorials
- Chimera removal - identifying sequences that are two-parent chimeras of more abundant output sequences

Sample metadata: formatting

[Keemei: cloud-based validation of tabular bioinformatics file formats in Google Sheets.](#)
Rideout JR, Chase JH, Bolyen E, Ackermann G, González A, Knight R, Caporaso JG. GigaScience. 2016;5:27.

https://keemei.qiime2.org



	A	B	C	D	E	F	G	H	I	J
1	#SampleID	BarcodeSequen	LinkerPrimerSeq	BodySite	Ye ar	Month?	Day	Subject	ReportedAntibioticUsage	DaysSinceExperimentStart
2	L1S8			ut	2008	10	28	1	Yes	0
3	L1S140			ut	2008	10	28	2	Yes	0
4	L1S57			ut	2009	1	20	1	No	84
5	L1S208			ut	2009	1	20	2	No	84
6	L1S76	ACTACGTGTGC	GTGCCAGCMG	gut	2009	2	17	1	No	112
7	L1S105	AGTGCGATGCC	GTGCCAGCMG	gut	2009	3	17	1	No	140
8	L1S257	CCGACTGAGAT	GTGCCAGCMG	gut	2009	3	17	2	No	140
9	L1S281	CCTCTCGTGAT	GTGCCAGCMG	gut	2009	4	14	2	No	168
10	L2S240	CATATCGCAGT	GTGCCAGCMG	left palm	2008	10	28	2	Yes	0
11	L2S155	ACGATGCGACC	GTGCCAGCMG	left palm	2009	1	20	1	No	84
12	L2S309	CGTGCATTATC	GTGCCAGCMG	left palm	2009	1	20	2	No	84

Head to tutorial and complete Section 1

Section 1: Importing, cleaning and quality control of the data

Taxonomic assignment of observed sequences (ASVs)

FeatureData[Sequence]

```
>feature5
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCCATGGTGAA
ATCCCTCGGCTCAACCGAGGAACTG
>feature4
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAA
AGGCTGGGGCTCAACCCCGGGACGG
>feature2
TACGTATGGGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGCGTAGGTGGTGGCTTAAGCGCAGGGTTTA
AGGCAATGGCTTAACTATTGTTCTC
>feature1
GACGGAGGATGCAAGTGTTATCCGGAATCACTGGGCGTAAAGCGTCTGTAGGTGGTTTACTAAGTCAACTGTTAA
ATCTTGAGGCTCAACCTCGAAATCG
>feature3
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGTAGGCGGTTAGGTAAGTCAGATGTGAA
AGCCCCGGGCTCCACCTGGGAATGG
```

Taxonomic assignment of observed sequences.

Reference Database
Silva, Greengenes, etc.

FeatureData[Sequence]

```
>feature5
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCCATGGTGAA
ATCCCTCGGCTCAACCGAGGAACTG
>feature4
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAA
AGGCTGGGGCTCAACCCCGGGACGG
>feature2
TACGTATGGGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGCGTAGGTGGTGGCTTAAGCGCAGGGTTTA
AGGCAATGGCTTAACTATTGTTCTC
>feature1
GACGGAGGATGCAAGTGTTATCCGGAATCACTGGGCGTAAAGCGTCTGTAGGTGGTTTACTAAGTCAACTGTTAA
ATCTTGAGGCTCAACCTCGAAATCG
>feature3
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGTAGGCGGTTAGGTAAGTCAGATGTGAA
AGCCCCGGGCTCCACCTGGGAATGG
```

FeatureData[Sequence]

```
>reference-sequence-1
TTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCAACATGGT
GACTCAACCGAGGAACTGAATTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTG
GCTTGGTAAGTCAACATGGTGACTCAACCGAGGAACTGAA
>reference-sequence-2
AACGTAGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAAAGG
```

FeatureData[Taxonomy]

reference-sequence-1	Bacteria; Proteobacteria; Gammaproteobact
reference-sequence-2	Bacteria; Bacteroidetes; Flavobacteria; F
reference-sequence-3	Bacteria; Proteobacteria; Deltaproteobact
reference-sequence-4	Archaea; Euryarchaeota; DSEG; 104A5

Taxonomic assignment of observed sequences.

Reference Database
Silva, Greengenes, etc.

FeatureData[Sequence]

```
>feature5
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCCATGGTGAA
ATCCCTCGGCTCAACCGAGGAACTG
>feature4
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAA
AGGCTGGGGCTCAACCCCGGGACGG
>feature2
TACGTATGGGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGCGTAGGTGGTGGCTTAAGCGCAGGGTTTA
AGGCAATGGCTTAACTATTGTTCTC
>feature1
GACGGAGGATGCAAGTGTTATCCGGAATCACTGGGCGTAAAGCGTCTGTAGGTGGTTTACTAAGTCAACTGTTAA
ATCTTGAGGCTCAACCTCGAAATCG
>feature3
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGTAGGCGGTTAGGTAAGTCAGATGTGAA
AGCCCCGGGCTCCACCTGGGAATGG
```

FeatureData[Sequence]

```
>reference-sequence-1
TTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCAACATGGT
GACTCAACCGAGGAACTGAATTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTG
GCTTGGTAAGTCAACATGGTGACTCAACCGAGGAACTGAA
>reference-sequence-2
AACGTAGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAAAGG
```

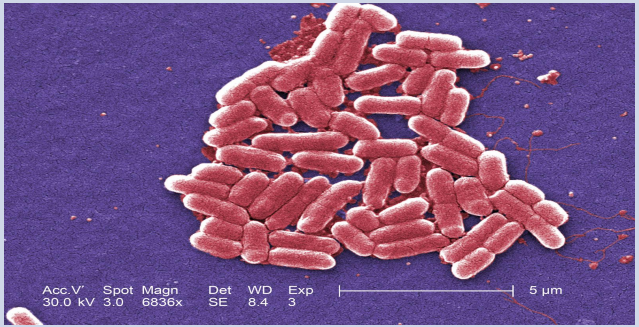
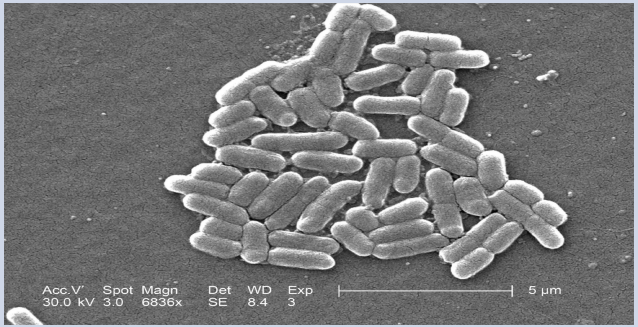
FeatureData[Taxonomy]

reference-sequence-1	Bacteria; Proteobacteria; Gammaproteobact
reference-sequence-2	Bacteria; Bacteroidetes; Flavobacteria; F
reference-sequence-3	Bacteria; Proteobacteria; Deltaproteobact
reference-sequence-4	Archaea; Euryarchaeota; DSEG; 104A5

Compare observed sequences to annotated reference sequences to make taxonomic assignments.

FeatureData[Taxonomy]

feature5	Bacteria; Proteobacteria
feature4	Bacteria; Proteobacteria
feature2	Bacteria; Bacteroidetes; Flavobacteria; Flavobacteriales
feature1	Bacteria; Proteobacteria
feature3	Bacteria; Proteobacteria; Deltaproteobacteria

	Ideal 16S	Real 16S
		
Kingdom	Bacteria	Bacteria
Phylum	Proteobacteria	Proteobacteria
Class	Gammaproteobacteria	Gammaproteobacteria
Order	Enterobacteriales	Enterobacteriales
Family	Enterobacteriaceae	Enterobacteriaceae
Genus	*Eschericia*	---
Species	*coli*	OTU 2445338
Strain	O157:H7	--

E. coli: http://media-3.web.britannica.com/eb-media//87/141087-050-24850517.jpg
http://upload.wikimedia.org/wikipedia/commons/d/d3/Staphylococcus_aureus_VISA_2.jpg

Classify Taxonomies

qiime2 feature-classifier (**Bokulich et al. 2018**)

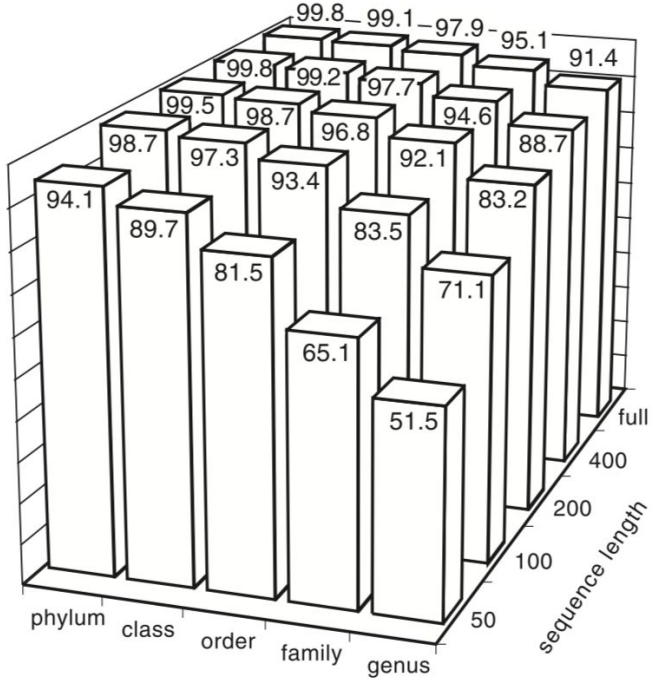


FIG. 1. Overall classification accuracy by query size (exhaustive leave-one-out testing using the Bergey corpus). Numbers are percentages of tests correctly classified.

Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Wang et al. **2007**. Applied and Environmental Microbiology.

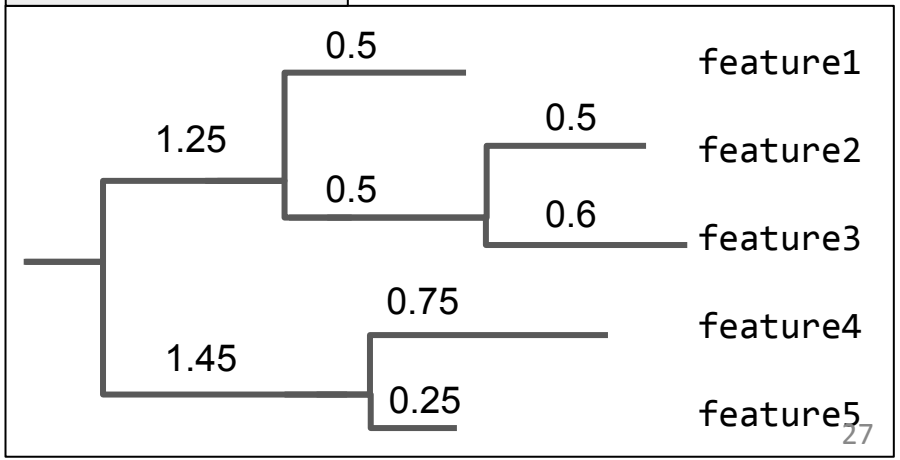
Phylogenetic reconstruction of observed sequences

FeatureData[Sequence]

```
>taxon5
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCCATGGTGAA
ATCCCTCGGCTCAACCGAGGAACTG
>taxon4
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAA
AGGCTGGGGCTCAACCCCGGGACGG
>taxon2
TACGTATGGGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGCGTAGGTGGTGGCTTAAGCGCAGGGTTTA
AGGCAATGGCTTAACTATTGTTCTC
>taxon1
GACGGAGGATGCAAGTGTTATCCGGAATCACTGGGCGTAAAGCGTCTGTAGGTGGTTTACTAAGTCAACTGTTAA
ATCTTGAGGCTCAACCTCGAAATCG
>taxon3
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGTAGGCGGTTAGGTAAGTCAGATGTGAA
AGCCCCGGGCTCCACCTGGGAATGG
```

Align sequences, filter highly variable (i.e., randomly evolving) positions, and build phylogenetic tree.

Phylogeny[Rooted]



Head to tutorial and complete Section 2&3

Section 2: Taxonomic Analysis

Section 3: Build a phylogenetic tree

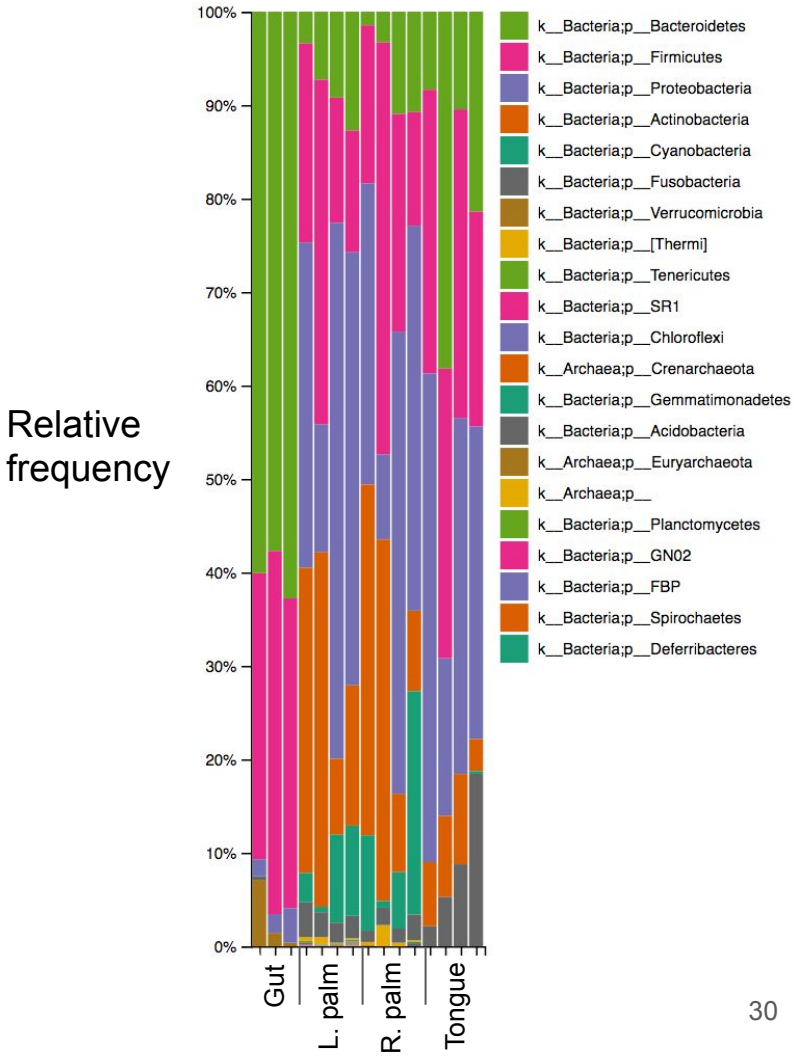
Basic visualizations and statistics

https://docs.qiime2.org/2021.4/tutorials/moving-pictures/#alpha-and-beta-diversity-analysis

Visualizing taxonomic profiles

Interactive barplots support:

- Taxonomic level selection
- Multi-level sorting
- Filtering
- Coloring
- Exporting plots (SVG) and raw data



emp-single-end-sequences
+ sample-metadata.tsv

sequences.fastq(.gz)

```
@HWI-6X_9267:1:1:25:1051
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGC
TTGGTA
+
abaaaa                                    ^\a^SYOZVVSV
YGYVDX
@HWI-6
TACGTA
GGCTTA
+
aa^^[
WWURZU
@HWI-6
TACGTA
TGGACA
+
aaab`a
[I^^aZ
@HWI-6
GACGGA
TTACTA
+
abaaaa
]]_Z_X
@HWI-6
TACGGA
TAGGTA
+
aaaba^
VH_PHO
```

barcodes.fastq(.gz)

```
@HWI-6X_9267:1:1:25:1051
AACGCAC
+
Bbbbbbb
@HWI-6X_9
AAGAGAT
+
bbbbbbb
@HWI-6X_9
AACGCAC
+
bbbbbbb
@HWI-6X_9
ACAGCAG
+
bbbbbbb
@HWI-6X_9
ACAGCTA
+
bbbbbbb
```

sample-metadata.tsv

SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

demux.qza

SampleData[SequencesWithQuality]

4ac2.fastq(.gz)

e375.fastq(.gz)

4gd8.fastq(.gz)

9872.fastq(.gz)

```
@HWI-6X_9267:1:1:25:1109
TACGGAGGGTGCGAGCGTTAATCGGAAT
TACTGGGCGTAAAGCGTACGTAGGCGGT
TAGGTAAGTCAGATGTGAAAGCCCCGGG
CTCCACCTGGGAATGG
+
aaaba^`a^N_`\_``a_a]Zaa^^\Z`
[M]a`[VYa^_X^_Z]NZ\`]TY\]_^R
VH_PHOWZM[PTRPTRYUBBBBBBBBBB
BBBBBBBBBBBBBBBB
```

table.qza

FeatureTable[Frequency]

	feature 1	feature 2	feature 3	Feature 4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

rep-seqs.qza

FeatureData[Sequence]

```
>feature5
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAG
TCCATGGTGAAATCCCTCGGCTCAACCGAGGAACTG

>feature4
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAG
TCTGATGTGAAAGGCTGGGGCTCAACCCCGGGACGG

>feature2
TACGTATGGGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGCGTAGGTGGTGGCTTAAG
CGCAGGGTTTAAGGCAATGGCTTAACTATTGTTCTC

>feature1
GACGGAGGATGCAAGTGTTATCCGGAATCACTGGGCGTAAAGCGTCTGTAGGTGGTTTACTAAG
TCAACTGTTAAATCTTGAGGCTCAACCTCGAAATCG

>feature3
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGTAGGCGGTTAGGTAAG
TCAGATGTGAAAGCCCCGGGCTCCACCTGGGAATGG
```

rooted-tree.qza

Phylogeny[Rooted]

Diversity and statistical analyses

sample-metadata.tsv

table.qza

FeatureTable[Frequency]

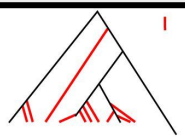
	feature 1	feature 2	feature 3	Feature 4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

sample-metadata.tsv

SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

rooted-tree.qza

Phylogeny[Rooted]



Diversity and statistical analyses

Comparing microbial communities

How many different ASVs are there? *Alpha diversity*
richness
evenness
both

How similar are pairs of samples? *Beta diversity*

Who is there? *Taxonomic profiling, differential abundance testing.*

Alpha diversity metrics operate on a single sample (i.e., within sample diversity).

Beta diversity metrics operate on a pair of samples (i.e., between sample diversity).

Does anything concern you about this table?

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature 5
4ac2	84	1	73	198	2
e375	24	2	44	176	1
4gd8	11	0	10	30	0
9872	0	0	25	2	0

Diversity metrics in ordinations are often impacted by the total frequency observed in samples, such that in this example `4gd8` might look more similar to `9872` than to `e375`.

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature 5
4ac2	84	1	73	198	2
e375	24	2	44	176	1
4gd8	11	0	10	30	0
9872	0	0	25	2	0

	Total frequency
4ac2	358
e375	247
4gd8	51
9872	27

This is most commonly handled by *rarefaction*, which is currently* a necessary evil. Frequencies are subsampled without replacement until all samples have the same total. Samples with fewer sequences than your *even sampling depth* will be filtered out of the feature table.

FeatureTable[Frequency]

	feature1	feature2	feature3	feature 4	feature5
g345	11	1	10	29	0
c5d7	4	0	7	40	0
f6ee	11	0	10	30	0
~~cfd3~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~

	Total frequency
g345	51
c5d7	51
f633	51
~~cfd3~~	~~0~~

* A good project would be developing diversity metrics that are not sensitive to total frequency.

Alpha rarefaction

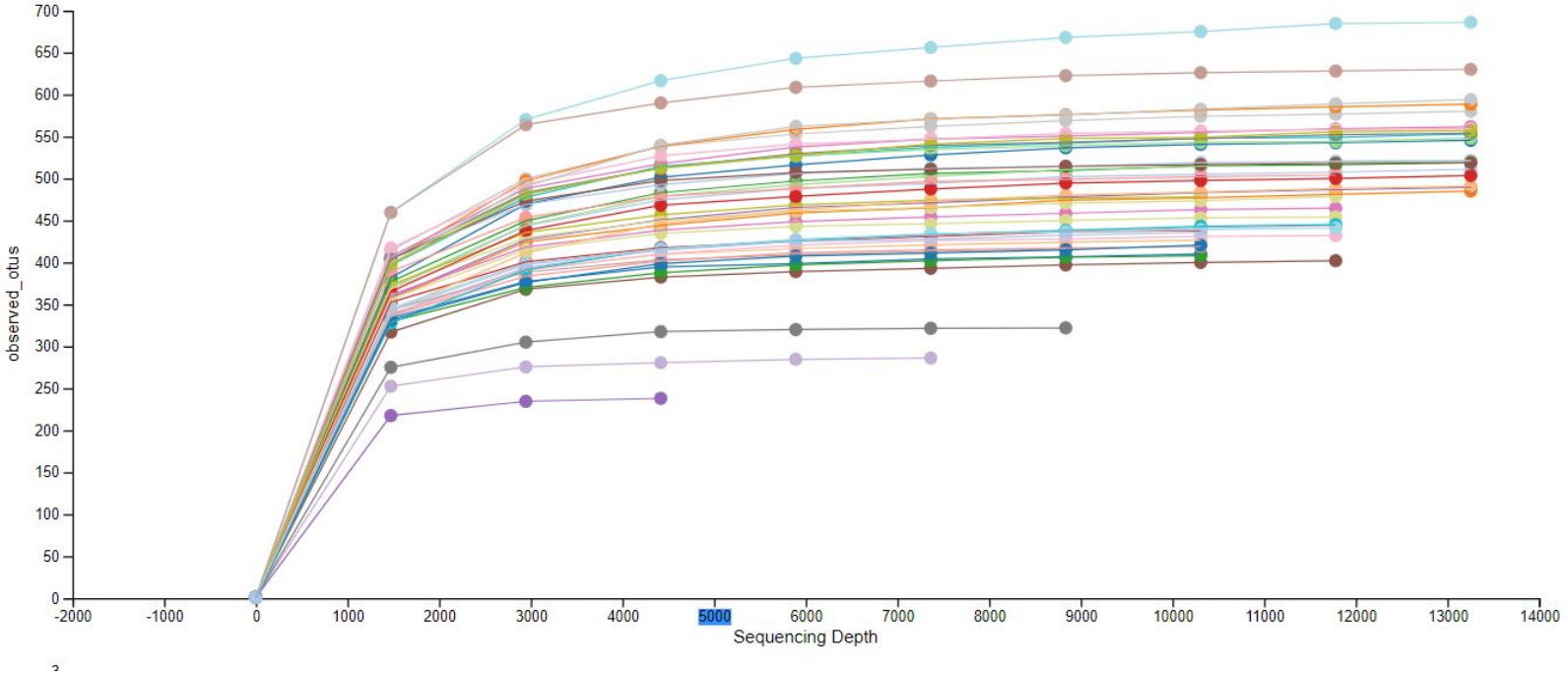
Download CSV

Metric

observed_otus

Sample Metadata Column

BarcodeSequence



Phylogenetic diversity metrics incorporate evolutionary relationships between taxa, but assume that we know what those relationships are. These require a phylogenetic tree.

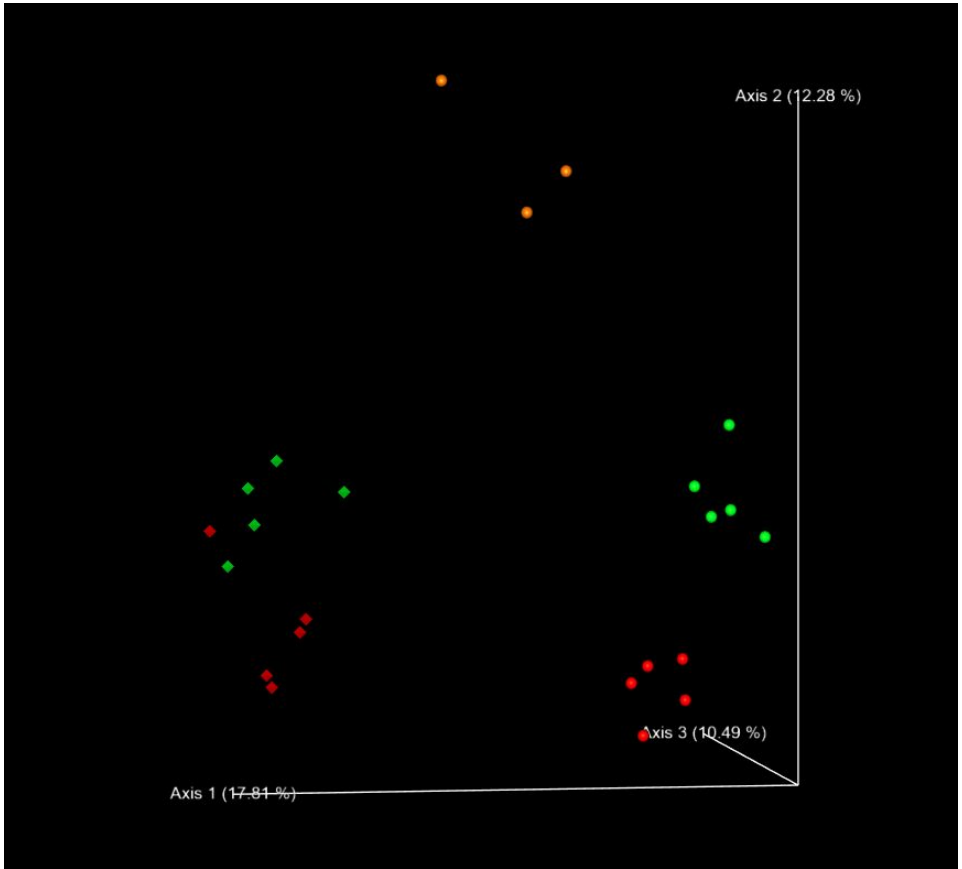
- Weighted Unifrac
- Unweighted Unifrac*

Non-phylogenetic diversity metrics assume that all taxa are equally related, so don't make assumptions about evolutionary relationships. No tree required.

- Bray-Curtis
- Jaccard*

*Unweighted doesn't consider abundance, just presence/absence

PCoA



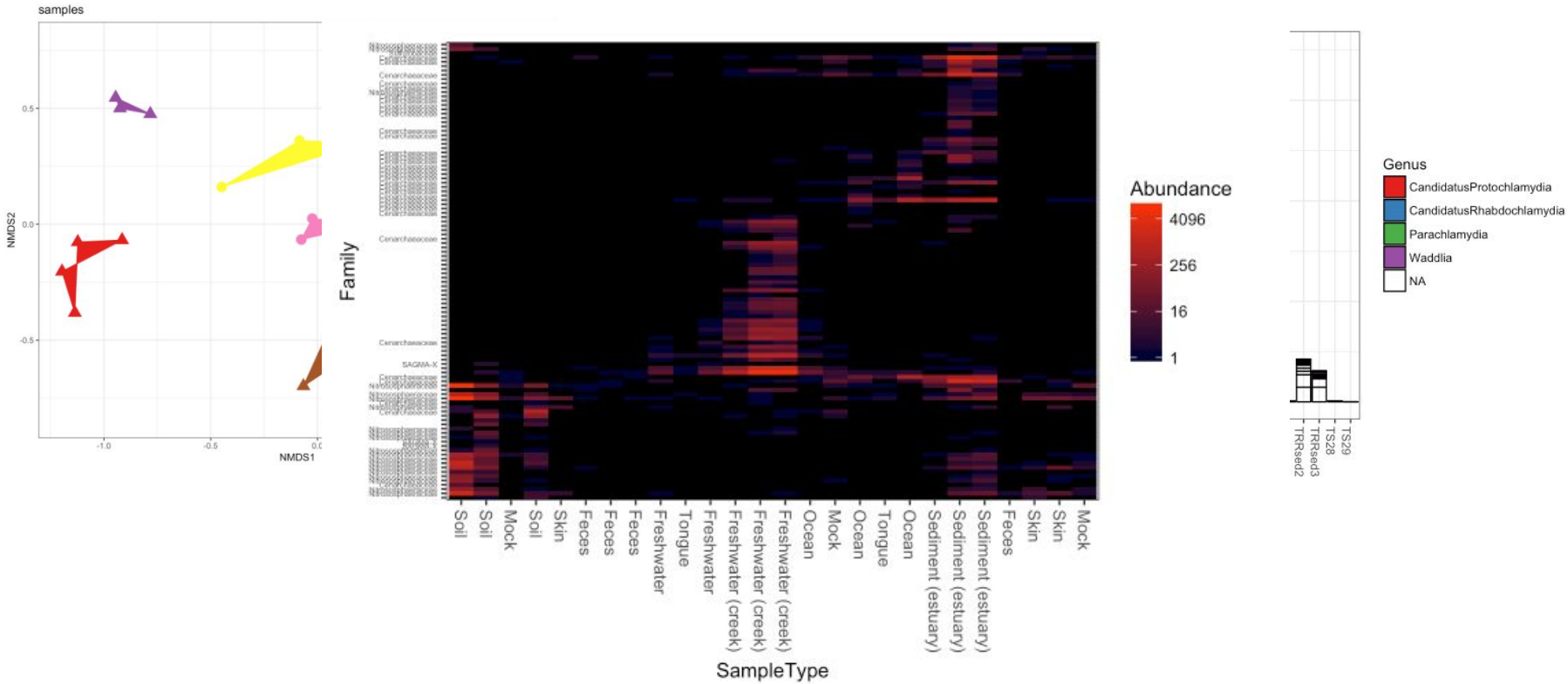
Color = Genotype

Shape = SW treatment

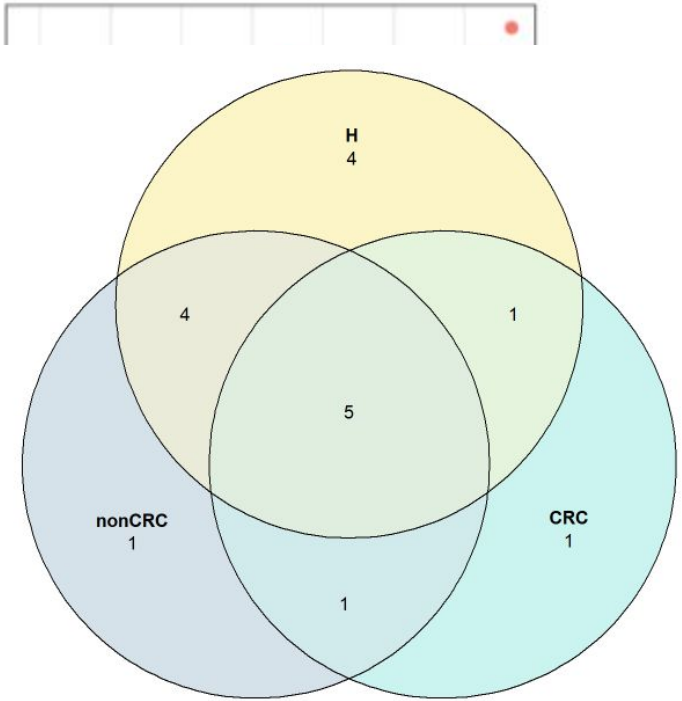
Head to tutorial and complete Section 4

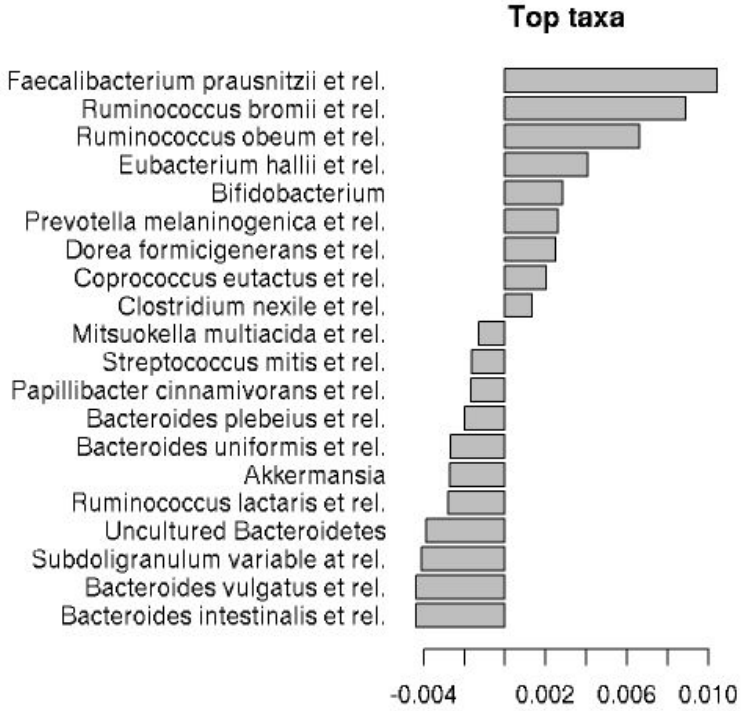
Section 4: Basic visualizations and statistics

QIIME2 → R → phyloseq



QIIME2 → R → microbiome





Other R packages

- indicspecies
- DeSeq2
- vegan
- MicrobiotaProcess
- metagenomeSeq
- mixOmics
- PICRUSt2
- LEfSe
- ALDEx2

Head to tutorial and complete Section 5

Section 5: Exporting data for further analysis in R